Open Data for Tourism: the case of Tourpedia

Angelica Lo Duca, Andrea Marchetti
Institute of Informatics and Telematics, National Research Council
via G. Moruzzi 1, 56124 Pisa (Italy)
email: [name].[surname]@iit.cnr.it

Abstract

Purpose

This paper describes Tourpedia, a Web site about tourism, built on open data provided by official government agencies. Tourpedia provides data under a public license.

Design/methodology/approach

Tourpedia is built upon a modular architecture, which allows a developer to add a new source of data easily. This is achieved through a simple mapping language, namely Tourpedia Mapping Language (TML), which maps the original open dataset model to the Tourpedia Data Model (TDM).

Findings

Tourpedia contains more than 70.000 accommodations, downloaded from open data provided by Italian, French and Spanish Regions.

Research limitations/implications

Tourpedia presents some limitations. Firstly, extracted data are not homogeneous and often they are incomplete or wrong. Secondly, Tourpedia contains only accommodations. Finally, at the moment Tourpedia covers only some Italian, French and Spanish Regions.

Practical implications

The most important implication of Tourpedia concerns the construction of a single access point for all Italian, French and Spanish open data about accommodations. In addition, a simple mechanism for the integration of new sources of open data is defined.

Social implications

The current version of Tourpedia opens also the road to three new possible social scenarios. Firstly, Tourpedia could be transformed into an open source of updated information about tourism. Secondly, Tourpedia could be empowered to support tours, which include some tourist attractions

and/or events and suggest the nearest accommodations. Finally, Tourpedia may help tourists to discover unknown places.

Originality

Tourpedia constitutes an access point for datasets providers, application developers and tourists, because it provides a unique Web site.

Keywords Tourism, Open Data, Web site, Public License.

1. Introduction

The tourism sector in Italy has always been one of the most important strategic. In 2017, Italian tourism has experienced a quantifiable improvement in the number of international arrivals: there was an increase of 4,5% visitors, from 116,944,243 in 2016 to 122,202,524 in 2017. Today Italy is the fifth most visited tourist destination in the world, with 50.7 million visitors, after France (84.5 million), the US (77.5 million), Spain (68.2 million) and China (56.9 million) (UNWTO 2017). Even from the point of view of employment, Italian tourism plays an important role. Thanks to the tourism sector, 2,609 million workers are employed, approximately 11.6% of the entire national employment.

Since peer-to-peer travel applications started to pervade the market, new tourists have been continuously attracted to Italian amenities (Moreno *et al.*, 2015). McGuire et al. demonstrated that consumers pay attention only to the place where the hotel is located (e.g. near some points of interest) and whether the hotel has or not a high assessment score (e.g. it is recommended by other travellers) (McGuire, 2015). In addition, Faizan demonstrated that the quality of a hotel web site plays an important role (Faizan, 2016).

Many Web sites offer a service where users can book and review accommodations. Among them, the most famous are: Booking.com[1], Trip Advisor[2], Expedia[3], Airbnb[4], Google Places[5] and Google Maps[6]. In addition to these Web sites, some social networks, such as Facebook[7] and Foursquare[8], contain all the information to enable tourists to organize their trips (Varkaris and Neuhofer, 2017). One of the major limitations of all these portals is that they do not release their data publicly. Only a

[1] http://www.booking.com
[2] https://www.tripadvisor.it/
[3] https://www.expedia.it/
[4] https://www.airbnb.it/
[5] https://developers.google.com/places/
[6] https://www.google.it/maps
[7] http://www.facebook.com
[8] http://foursquare.com

small part of their data can be accessed, sometimes through Web APIs, which are provided with some restrictions. This constitutes a strong limitation for tourism communities that would like to build new services exploiting these data.

This paper describes Tourpedia[9], a Web site for tourism, which exploits official open data provided by government agencies. The main objective of Tourpedia consists in defining, designing and implementing new technologies, which give a common structure to open data released by government agencies. As use case of Tourpedia, official open datasets about accommodations provided by Italian, French and Spanish Regions are exploited. The idea is to extend Tourpedia also with other tourism resources, such as attractions, events and restaurants.

Open data about tourism are often distributed through different Web sites and in different formats or data structures. Tourpedia aims at unifying all these open data in order to provide a single Web site to access open data about tourism. The system architecture behind Tourpedia is composed of different modules, which allow a developer to add a new data source easily and without compromising the already imported sources. Tourpedia provides also a mechanism to map the original schema defined by a data source to the Tourpedia Data Model (TDM), through a simple mapping language, called Tourpedia Mapping Language (TML). The TDM is very generic and can be extended easily to add new features. Currently Tourpedia contains more than 70.000 accommodations, collected by 21 of the official open data Web sites provided by Italian, French and Spanish Regions (12 Web sites are from Italy, 6 from France and 3 from Spain). All the available data are aggregated, updated continuously, stored in a local database, released under a public license and can be accessed through a Web API and a Web application.

All the concepts and methodologies adopted in Tourpedia are described in the remainder of the paper. Section 2 analyses related work, while Section 3 describes Italian, French and Spanish open data about accommodations. Section 4 illustrates the previous version of Tourpedia (namely Tourpedia 1.0) and in Section 5 the current version of Tourpedia. Finally, in Section 6 conclusions and future work are given.

2. Related Work

The use of open data in different fields opens many business possibilities, such as the reduction of application costs as well as the combination of multiple sources of data (Garcia *et al.*, 2015). However, possibilities of open data in the tourism sector are still unexplored (Pesonen and Lampi, 2016, Li *et al.*, 2018). In fact the literature about this topic is relatively reduced. McNaughton et al. (McNaughton *et al.*, 2016) investigate the topic of tourism and open data in Jamaica, while Urata et al. (Urata *et al.*, 2016) propose a mechanism to promote the construction of open data about sightseeing events in Japan, as well as an app to access them.

Although the association between open data and tourism is very recent, there are many projects similar to Tourpedia, which exploit more consolidated sources of data, such as user generated

[9] http://tour-pedia.org/it

content and social media. Scarinci et al. (Scarinci et Myers, 2014) describe a semantic Web framework, which aggregates some practice standards for the creation of lodging properties. Soualah-Alila et al. (Soualah-Alila *et al.*, 2016) illustrate DataTourism, a tool for the management of heterogeneous data related to tourism and derived from different sources.

TouriNet[10] is a project, which aims at developing new technologies to improve tourism businesses, enhancing their reputation on the Web and promoting the creation of synergies between different activities. The TouriNet project exploits sources such as TripAdvisor and Booking.com, which do not provide data under public license terms.

Many different Web sites, offering tourism services, exist in the literature. They can be classified in different categories: a) open Web sites b) proprietary Web sites c) social networks. Among the most important open Web sites, there are OpenStreetMap[11] and Wikidata[12]. OpenStreetMap is a world map created through crowdsourcing. It contains many accommodations, attractions and restaurants. Unfortunately, information concerning the Italian tourism is not complete. Wikidata[13] is a free, collaborative and multilingual database, which can be edited by anyone. From the tourism point of view, Wikidata does not provide much information.

Among the proprietary Web sites, the most important are Trip Advisor and Google Maps[14]. Trip Advisor is a Web site created for tourists. It contains data about hotels, attractions and restaurants. For each of them, the user can write a review and give a rating. Google Maps is not a portal for online booking, but a Web application that provides the ability to show data on a map. Available information is very detailed and each hotel is connected to other portals such as Booking.com.

Social media are exploited by businesses to advertise their activities (Zeng and Gerritsen, 2014, Mariani *et al.*, 2016). Among them, the most popular ones for tourism are Facebook, Foursquare and Google Places[15]. Facebook is a very popular social network, which contains also several pages dedicated to accommodation. However, since tourism is not the main objective of Facebook, page search is not easy. Foursquare is a social media, which allows users to search for information around their current position, such as shops, businesses and places of artistic or cultural interest. Foursquare contains many registered sites, but in terms of accommodation facilities it is lacking. For example, Foursquare contains only 90 registered hotels located in Florence, in contrast with Booking.com, which contains 384 hotels located in the same town.

Google Places[16] allows Google users to introduce their business. Thanks to it, users can create a page related to their business, including photos, location, opening hours, closing days, and so on. Users can leave reviews and rate their level of satisfaction.

[10] http://www.tourinet.it/
[11] https://www.openstreetmap.org/
[12] https://www.wikidata.org/
[13] https://www.wikidata.org/
[14] https://www.google.it/maps
[15] https://developers.google.com/places/
[16] https://developers.google.com/places/

3. Open Data about Accommodations

In 2005 the Open Knowledge Foundation[17] defined open data as *data that can be freely used, shared and built on by anyone, anywhere, for any purpose*. Open data are published mainly by government agencies and public administrations, which collect their data manually from their data portals. Then they aggregate data through a semi-automatic procedure in order to produce datasets. A dataset is a collection of information related to a specific category, collected over a specific period of time. Open data can be used for different purposes, ranging from the guarantee of democratic theory (Amichai-Hamburger *et al.*, 2008), to the vitality of civic society (Bertot *et al.*, 2010). Recently, research on open data has been exploring open data as a catalyser of innovation (Lakomaa and Kallberg, 2013, Maccani *et al.*, 2015). In addition, open data can be used to enrich existing online services and offerings. Without loosing in generality, this paper focuses on open data about tourism. Tourism is considered the first industry where open data is applied (Longhi *et al.*, 2014, Wu *et al.*, 2014, Pantano *et al.*, 2017).

Without loosing in generality, this paper pays attention to accommodations provided by Italian, French and Spanish Regions, but the same analysis can be done also for other tourism sectors, such as restaurants and events, and for other Countries. Regarding Italy, most of the twenty Italian Regions provide an access point for open data: only Marche and Sicily do not. Unfortunately, among the other eighteen Italian Regions, only 12 provide a dataset about accommodations. Furthermore, some datasets are limited: Basilicata provides accommodations only for the province of Matera, while the portal of Trentino-Alto Adige only those of the province independent of Trento. Finally, Sardinia provides a dataset with a format not readable by machines (PDF) thus it is discarded from this analysis. Table 3 summarizes the results of this first phase.

Region	Open Data Portal	Availability of Dataset	Format
Abruzzo	opendata.regione.abruzzo.it	No	-
Basilicata	www.aptbasilicata.it	Yes	.xsl
Calabria	dati.reggiocal.it	No	-
Campania	opendatacampania.it	No	-
Emilia-Romagna	dati.emilia-romagna.it	Yes	.csv
Friuli-Venezia Giulia	dati.friuliveneziagiulia.it	Yes	.csv

[17] http://okfn.org

Lazio	dati.lazio.it, dati.comune.roma.it	Only Rome Municipality	.csv
Liguria	www.regione.liguria.it	Yes	.csv
Lombardia	dati.lombardia.it	Yes	.csv
Marche	goodpa.regione.marche.it	Yes	.csv
Molise	n / a	-	-
Piedmont	dati.piemonte.it	Yes	.csv
Puglia	dati.puglia.it	Yes	.csv
Sardinia	opendata.sardegna.it	Yes	.pdf *
Sicilia	n / a	-	-
Tuscany	dati.toscana.it	Yes	.csv
Trentino-Alto Adige	dati.trentino.it	Only prov. Trento	.xml
Umbria	dati.umbria.it	Yes	.csv
Aosta	www.regione.vda.it	No	-
Veneto	dati.veneto.it	Yes	.csv

*Discarded

Table 1. The table shows the results of the first phase of open data search about accommodations. For each Italian Region, its official link to open data is provided, as well as the availability of data about accommodations and data format.

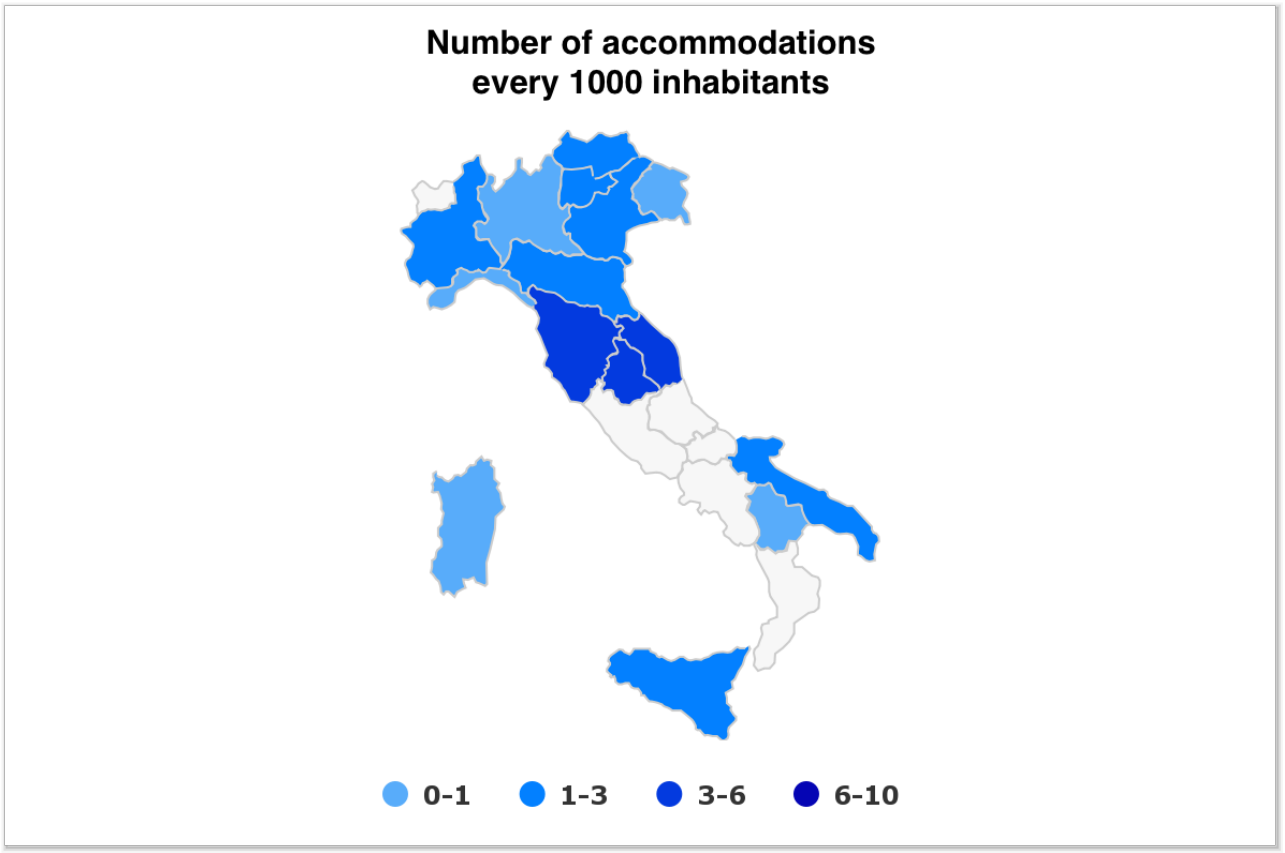
Number of accommodations
every 1000 inhabitants
0-1
1-3
3-6
6-10

Fig. 1

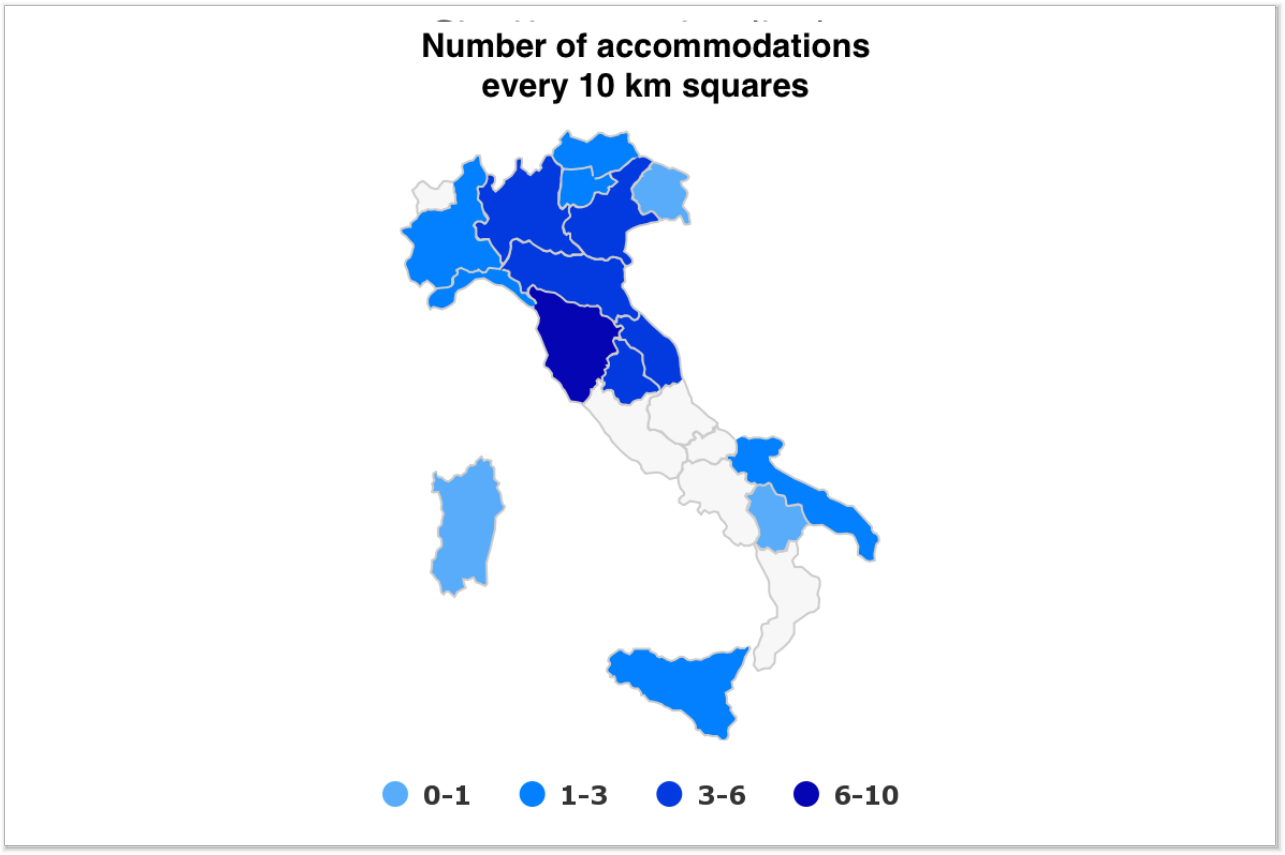

Fig. 2

Fig. 1 and Fig. 2 show accommodations density, normalized for every 1,000 inhabitants and for every 10 km^2, respectively. According to Figure 1, Regions located in the central part of Italy provide datasets with the highest density (between 3 and 6 accommodations for every 1,000 inhabitants. Other Regions, such as Lombardy, Liguria and Friuli-Venezia Giulia have a low density of accommodations (between 0 and 1). This could be interpreted in two ways: 1) the dataset is not complete, or 2) in these Regions effectively there are few accommodations. According to Figure 2, the situation changes slightly: Tuscany remains the Region with the highest density (between 6 and 10 accommodations for every 10 km^2), in contrast to Umbria and Marche that descend into the lower range (between 3 and 6).

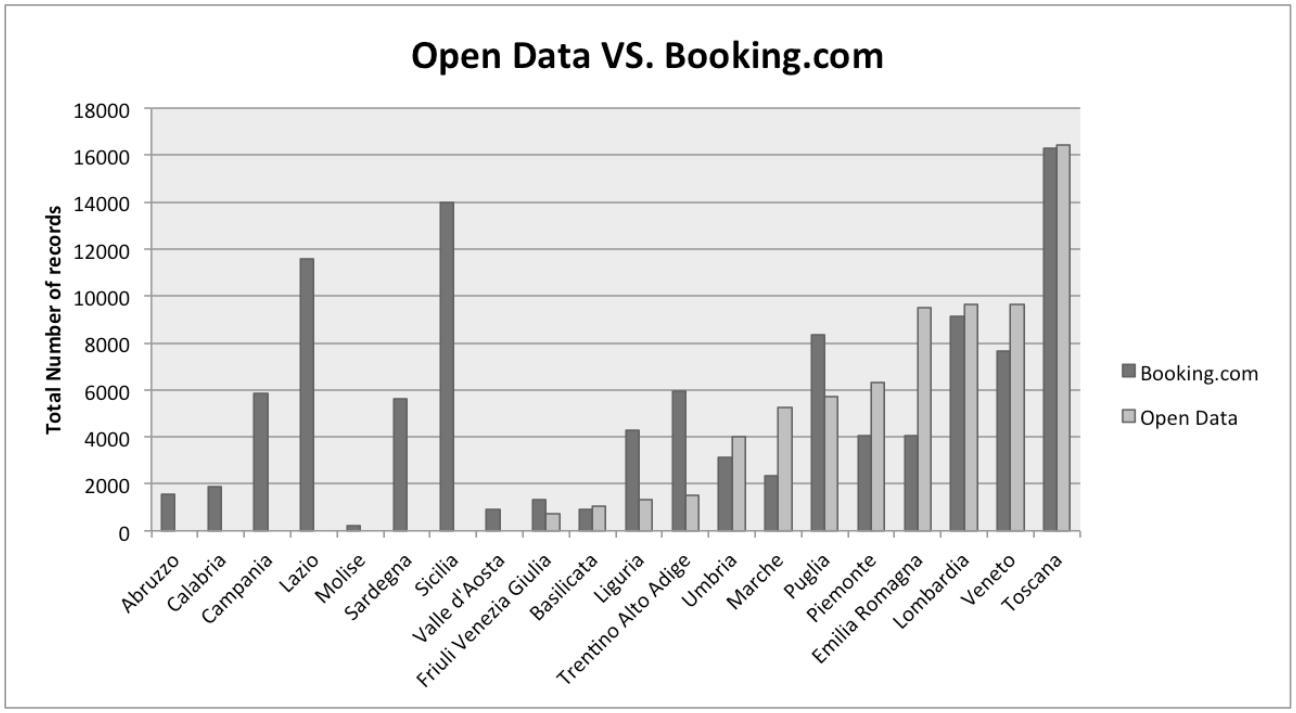

Fig. 3

In order to analyse the completeness of the various datasets, a comparison is done with data extracted from the Booking.com portal (Fig. 3): for each Region, the total number of accommodations is extracted from Booking.com. Eight open data contain much information about accommodations than the famous portal. At the same time, Friuli-Venezia Giulia, Liguria, Puglia and Trentino-Alto Adige seem to show a significant gap between the two: statistics reveal an inadequacy of information in open data.

3.1 Dissemination of results

In order to make people and Regions aware about problems related to open data and tourism, a Web site was built[18]. This portal contains all the described statistics about open data, as well as a geographical map with all the collected accommodations.

[18] http://www.tour-pedia.org/it

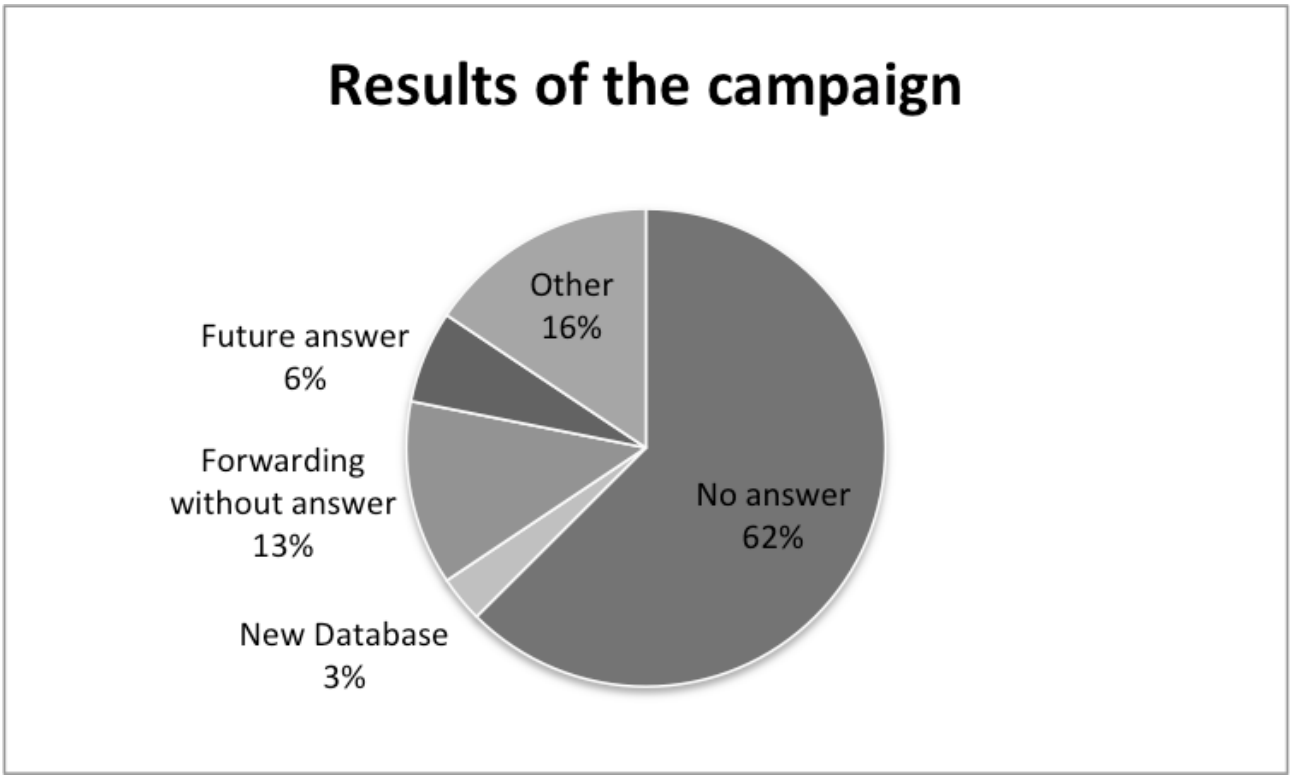

Fig. 4

In addition, a campaign was done to attract interest of Regions about this topic. The campaign consisted in contacting all Regions and people responsible for open data by email. The campaign aimed at searching a partnership that would allow obtaining more updated information or, where they are lacking completely, a portal for data download. The total number of sent emails was thirty-two, but the answers were only twelve. Fig. 4 summarizes results: the 62.5% of emails did not receive any answer; the 12.5% were only forwarding notifications for the following competence office, but they never got further response; the 6.25% declared that they would have responded in the next days, but they never did. The only Region, which answered to the sent emails, was Basilicata. They provided us with another URL, which contains a more complete dataset.

3.2 Other Countries

A similar study was done also for other two European Countries, France and Spain, although, emails were sent only to the Italian Regions. Table 2 and Table 3 show the results of the study for France (Overseas departments were not included in this study) and Spain, respectively. Regarding the data format, the CSV format was preferred, thus if available, the other possible formats are not specified in the table. Tourpedia includes also accommodations about these two countries.

Region	Open Data Portal	Availability of Dataset	Format
Grand Est	http://data.grandest.fr/	No	-
Nouvelle-Aquitaine	-	-	-
Auvergne-Rhône-Alpes	http://opendata.auvergnerhonealpes.eu/	Only for the Cantal Department	csv
Bourgogne-Franche-Comté	https://www.ideobfc.fr/accueil	Yes	different formats, included csv
Bretagne	http://www.data-tourisme-bretagne.com	Yes	csv
Centre-Val de Loire	http://sig-crcentre.opendata.arcgis.com/	No	-
Corse	http://www.opendata.corsica/	No	-
Île-de-France	http://data.iledefrance.fr	Yes	csv
Occitanie	http://www.opendatalab.fr/	No	-
Hauts-de-France	https://opendata.hautsdefrance.fr/	No	-
Normandie	http://www.opendata-27-76.fr/	No	-
Pays de la Loire	http://data.paysdelaloire.fr/	Yes	Different formats, included csv
Provence-Alpes-Côte d'Azur	http://opendata.regionpaca.fr/	Yes	xsl, ods

Table 2. Open data about accommodation in France.

Region	Open Data Portal	Availability of Dataset	Format

Andalucía	http://www.juntadeandalucia.es/	No	-
Aragón	http://opendata.aragon.es/	Yes	px
Principado de Asturias	https://goo.gl/RvRf1A	No	-
Islas Baleares	http://www.caib.es/caibdatafront/catalog?lang=es	No	-
Canaries	http://opendata.gobiernodecanarias.org/opendata/inicio/index.html	No	-
Cantabria	http://www.icane.es/linked-open-data	No	-
Castilla-La Mancha	http://datosabiertos.castillalamancha.es/	Yes	html
Castilla y León	https://datosabiertos.jcyl.es/	Yes, but the link does not work	csv
Cataluña	http://governobert.gencat.cat/ca/dades_obertes/	Yes	xml
Comunidad Valenciana	http://www.dadesobertes.gva.es/	Yes	kml
Extremadura	-	-	-
Galicia	http://abertos.xunta.gal/	The portal does not work	-
La Rioja	http://www.larioja.org/dato-abierto-rioja/es?	No	-
Comunidad de Madrid	http://datos.madrid.es/	Yes, but only for the city of Madrid	xml, rdf
Región de Murcia	http://datosabiertos.regiondemurcia.es/	Yes	csv
Comunidad Foral de Navarra	https://gobiernoabierto.navarra.es/es/open-data	Yes	csv

País Vasco	http://opendata.euskadi.eus/inicio/	Yes	xml, json, geojson, …

Table 3. Open data about accommodation in Spain.

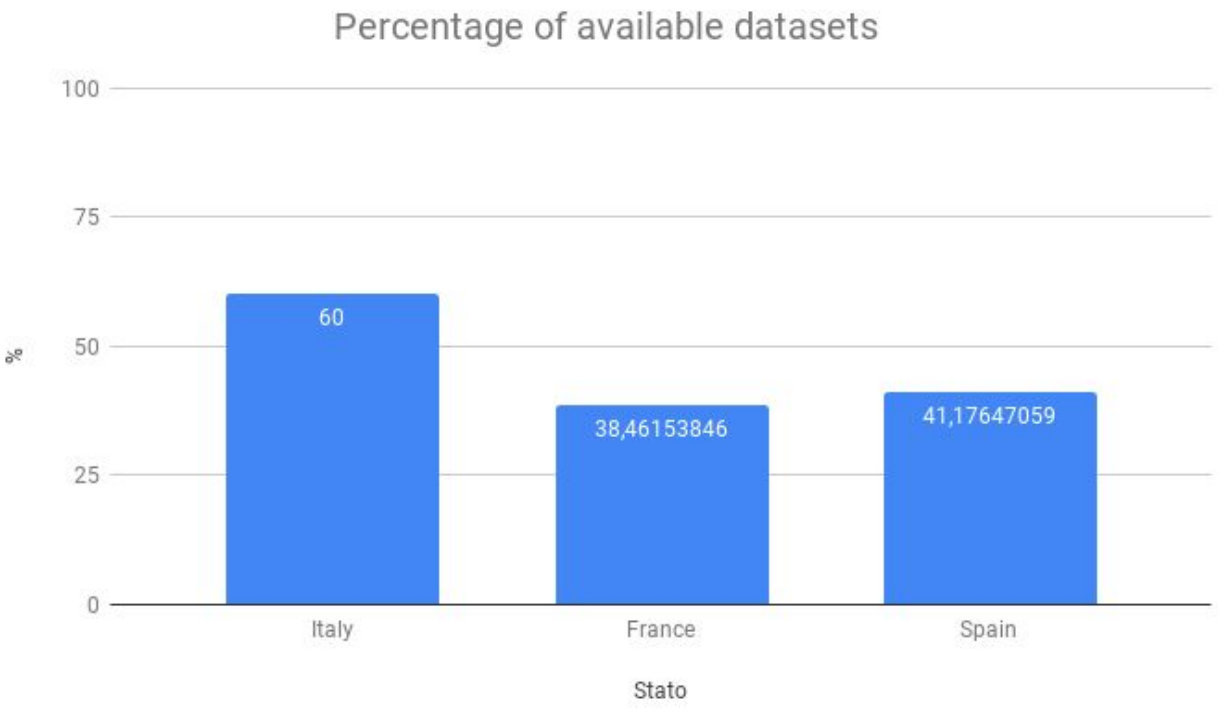

Fig. 5

Fig. 5 illustrates the percentage of available accommodation datasets for every country included in this study. The percentage has been calculated as the ratio between the number of available datasets (excluded not complete ones) and the number of Regions of the considered country. It is interesting to note how Italy reaches the highest percentage.

4. Tourpedia 1.0

Tourpedia is a project aiming at unifying all the open data about tourism in order to build a reference service for tourists. The idea of Tourpedia was generated within the OpeNER project[19], a FP7 European project, whose main goal was to build a set of tools for Named Entity Recognition and Sentiment Analysis. The role of Tourpedia within OpeNER was to provide a repository of

[19] http://opener-project.eu

tourism entities to support Named Entity Disambiguation. Thus, initially, Tourpedia (Cresci *et al.*, 2014, Gazzè *et al.*, 2015) was thought as an encyclopaedia of tourism, just like Wikipedia is an encyclopaedia for a general domain. In its original version, Tourpedia was based on data extracted from four social media: Booking.com, Facebook, Foursquare and Google Places and contained entities related to four categories of tourism places: accommodations, attractions, points of interest and restaurants. Data were collected for seven European cities: Amsterdam, Dubai, Paris, Rome, Barcelona, London and Berlin. Data about Tuscany were also collected.

Recently, the original project, based on data extracted to social media, changed its focus to open data, because open data can be reused and redistributed, under less restrictive licenses than social media. In fact, in most cases, data extracted from social media cannot be stored permanently and cannot be redistributed to third parties. For this reason, Tourpedia has moved to open data, which are more flexible. In contrast, open data can be easily modified and distributed to third parties, provided that license terms are respected. Thus, the main difference between Tourpedia 1.0 and Tourpedia 2.0 consists in the sources of data and the way they are collected: Tourpedia 1.0 extracted data from social media through ad-hoc crawlers, while Tourpedia 2.0 extracts data from open data through a more generic mechanism, which can be easily adapted to many sources.

5. Tourpedia 2.0

Tourpedia is based on a modular architecture, which permits a developer to add new open datasets easily and keep them updated. Fig. 6 illustrates the Tourpedia architecture. Starting from the bottom of the figure, the system is composed of the following modules:

1) *Sources* - for each source of data, a new module is built. This module downloads the specific dataset from a given source. In the specific case of accommodations, sources are represented by Italian, French and Spanish Regions. However, new sources could be added dynamically to the system, without compromising the existing ones. For each source, the URL of the dataset must be specified, as well as its mapping file, as described later;
2) *Mapping* - each dataset is mapped to a common mapping schema, as defined by the Tourpedia Data Model;
3) *Storage* - each mapped dataset is saved into a local no-relational database;
4) *Web API* - this module allows a user to access data directly.;
5) *Web application* - at the top of the stack, a Web application can be built. An example of Web application could be a geographical map, with all the accommodations related to a Region.

As additional services, Tourpedia provides also a mechanism to update datasets periodically (i.e. once a day) and to log extracted information, such as the last modified date associated to each dataset and the number of downloaded data. Simply scheduling the crawler for each source every day, through a cron job, does datasets update.

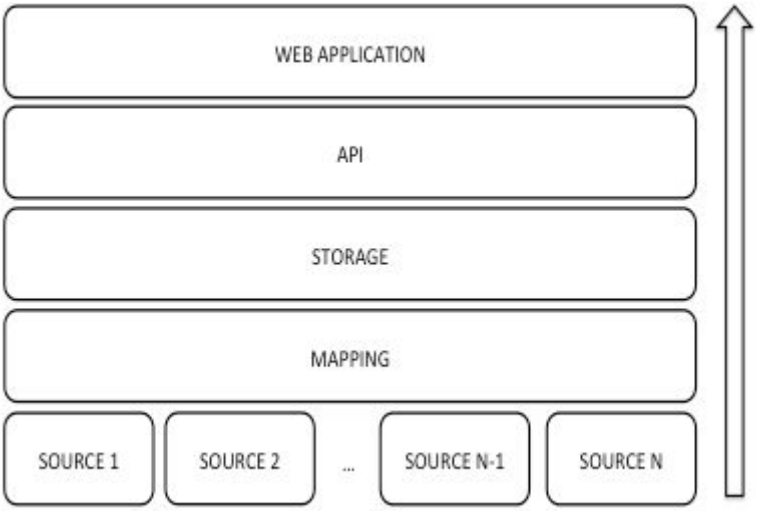

Fig. 6

5.1 Sources

Starting from the bottom of the Fig. 6, there are N sources of open data, which provide their data in their portals. Each source provides its data in a specific format and schema and datasets distributed by different sources have a different schema. The schema implies the list of fields described by each dataset. A specific crawler must be implemented to download the dataset and update it. Tourpedia defines a template of crawler, which can be personalized for each source.

Tourpedia maintains a configuration file, where all the sources must be specified. For each source, the URL to the dataset, the dataset format and the mapping file must be defined, as follows:

```
[MySource]
url = http://path/to/dataset.csv
format = CSV
separator = |
encoding = utf8
mapping[name] = 1
...
```

In square brackets, the name of the source is indicated (MySource in the example). The keyword *url* specifies the url to the dastaset, while the keyword *format* is optional and specifies the dataset format. If the dataset format is .csv, it is possible to specify also the separator character (e.g. the pipe) and the encoding. Finally, the mapping structure must be given, as described in the next paragraph.

5.2 Mapping

Each downloaded dataset provides a specific data model. Tuscany has the simplest data model, with only 14 fields, while Lombardia defines the most complicated data model, with 35 fields. In order to standardize all the datasets, each of them is converted to a specific data model, which is defined by Tourpedia, i.e. the Tourpedia Data Model (TDM). This allows the system to standardize all data and to have a single data model for all the sources. Many different ontologies have been proposed in the literature to represent data related to tourism (Prantner *et al.*, 2007): the Hi-Touch ontology (Legrand, 2004), the Harmonize Ontology (Dell'Erba *et al.*, 2002), the Acco ontology (Hepp, 2013), the Hontology ontology (Chaves *et al.*, 2012) and the taxonomy of hotel features proposed by Gibbs et al. (Gibbs *et al.*, 2016). Currently the TDM does not exploit anyone of the existing ontologies, because it takes simply all the possible fields from the open datasets and tries to unify them. The next step will be to represent data in one or more existing data models.

The TDM contains all the possible fields of all the open datasets and renames them to a single tag in English. Table 4 shows the structure of the TDM.

Field	**Type**	**Description**	**Nr. of datasets containing the field**
_id	string	ID	12
name	string	Accommodation name	12
description	string	Description of the accommodation	12
category	string	Accommodation category (e.g. Bed and Breakfast, hotel, ..)	1
address	string	Address where the accommodation is located	12
postal-code	string	Postal code associated to the address	12

city	string	City where the accommodation is located	12
province	string	Acronym of the province	11
hamlet	string	Hamlet associated to the address	6
locality	string	Locality associated to the address	6
region	string	Region associated to the address	12
latitude	double	Latitude where the accommodation is located	5
longitude	double	Longitude where the accommodation is located	5
number of stars	integer	Number of stars	10
telephone	string	Telephone number	12
telephone2	string	Alternative telephone number	2
cellular phone	string	Cellular number	3
fax	string	Fax number	11
web site	string	Link to the Web site	11
email	string	Email address	12
beds	integer	Number of beds	8
rooms	integer	Number of rooms	7

suites	integer	Number of suites	1
toilets	integer	Number of toilets	4
facilities	list	List of facilities	4
sports equipment	list	List of sports equipment	3
languages	list	List of spoken languages	3
breakfast	int	Specify whether or not breakfast if available. Possible ranges are 0 and 1	1
congress halls	string	Available congress halls	1
opening period	string	Opening period	3
credit/debit card	list	List of support credit/debit cards	2
location	list	List of locations (e.g near the airport, near the train station,…)	4
manager	string	The name of the accommodation's manager	1
elevation	integer	The accommodation elevation	1
high season price	list	A list containing the high season price for each category of room	1
low season	list	A list containing the low season price for each	1

price		category of room	
photo	list	A list of photos URLs about the accommodation	1
chain	string	The chain which the accommodation belongs to	2

Table 4. The Tourpedia Data Model.

addr	addr_number	acco_name	beds (divided by 100)
Via Moruzzi	3	Hotel Bologna	3000
Via Mancini	23	Hotel Roma	4200

Table 5. Example of CSV file.

For sources exporting datasets in CSV, Tourpedia defines a mapping language, namely Tourpedia Mapping Language (TML) to specify a mapping strategy from the source data model to the TDM. For sources exporting datasets in the other formats, firstly a conversion to CSV is done. At the moment, only the conversion from XSL and ZIP to CSV is done, but in the future Tourpedia will support also the conversion from other formats. The TML allows also defining some other basic operations, such as the aggregation of two fields or the normalization of a number. As example, consider the mapping file associated to the CSV file of Table 5 is the following:

```
address = 0,1
name = 2
beds = 3/100
```

The field *address* of the TDM is given by the concatenation of columns 0 and 1 in the CSV file (columns start from 0), while the column 2 gives the field name. Column 3 provides the *beds* field, which must be divided by 100 (as indicated by the title of the column in the CSV). In this case, the operator / is used. The TML allows defining also other operations, which must be specified by the < operator:

```
name = 2<utf8
province = 2<province
```

In this case, the column 2, converted in utf8, gives the field name. The column 2 gives the field province and then the acronym of the province is calculated. This case applies when the province value is provided as full name and not as an acronym. Other operations following the < operator can be defined easily by the user.

5.3 Dataset Enrichment

Some datasets show some deficiencies in their schema, such as geographical coordinates. Of the twelve Italian open data at our disposal, only five (Emilia-Romagna, Lombardia, Marche, Puglia and Tuscany) contain geographical coordinates. In order to provide also the remaining open data with geographical coordinates, Tourpedia implements a geocoding system, based on the Google Geocoding API[20] that receives the address as input and returns the geographical coordinates as output.

At the moment, only missing geographical coordinates have been added to Tourpedia. As future work, it could be important to integrate also other information, such as offered services and facilities, as well as the link to the associated social network. To do this, one possible solution could be to take advantage of the collaborations established with the Regions or contact the accommodation owners directly, since some of them also provided an email address in the open dataset.

5.4 Storage, API and Web Application

Data are stored into a no-relational database, i.e. Mongo DB[21]. The integrated dataset can be queried through a Web API[22], which receives the following parameters as input:

- *category* (mandatory) - the tourism category. At the moment, the only acceptable value is accommodation;
- *region* (optional) - the Region where to search for accommodations;
- *country* (optional) - the Country where to search for accommodations;
- *min_beds, max_beds* (optional) - the minimum and maximum number of beds which the accommodation must have;
- *min_latitude, max_latitude* (optional) - the minimum and maximum latitude where the accommodation must be located;
- *min_longitude, max_longitude* (optional) - the minimum and maximum longitude where the accommodation must be located.

In addition, each field defined in the TDM could be added as a search parameter. An example of query is the following:

http://tour-pedia.org/it/api/query.php?category=accommodation&region=Basilicata&min_beds=100

[20] https://developers.google.com/maps/documentation/geocoding/start
[21] https://www.mongodb.com
[22] http://tour-pedia.org/it/api/query.php

This query asks the Web API for all the accommodations located in Basilicata and having a minimum number of beds equal to 100.

The last layer of the Tourpedia architecture is represented by a Web application. Anyone can define his/her own application, on the basis of the provided Web APIs. As example, a basic Web application has been implemented. Such an application represents all the accommodations in a geographical map. In order to make the Web application user-friendly, the system implements a layout similar to that of Google Maps[23]. Fig. 7 shows a snapshot of the Web application.

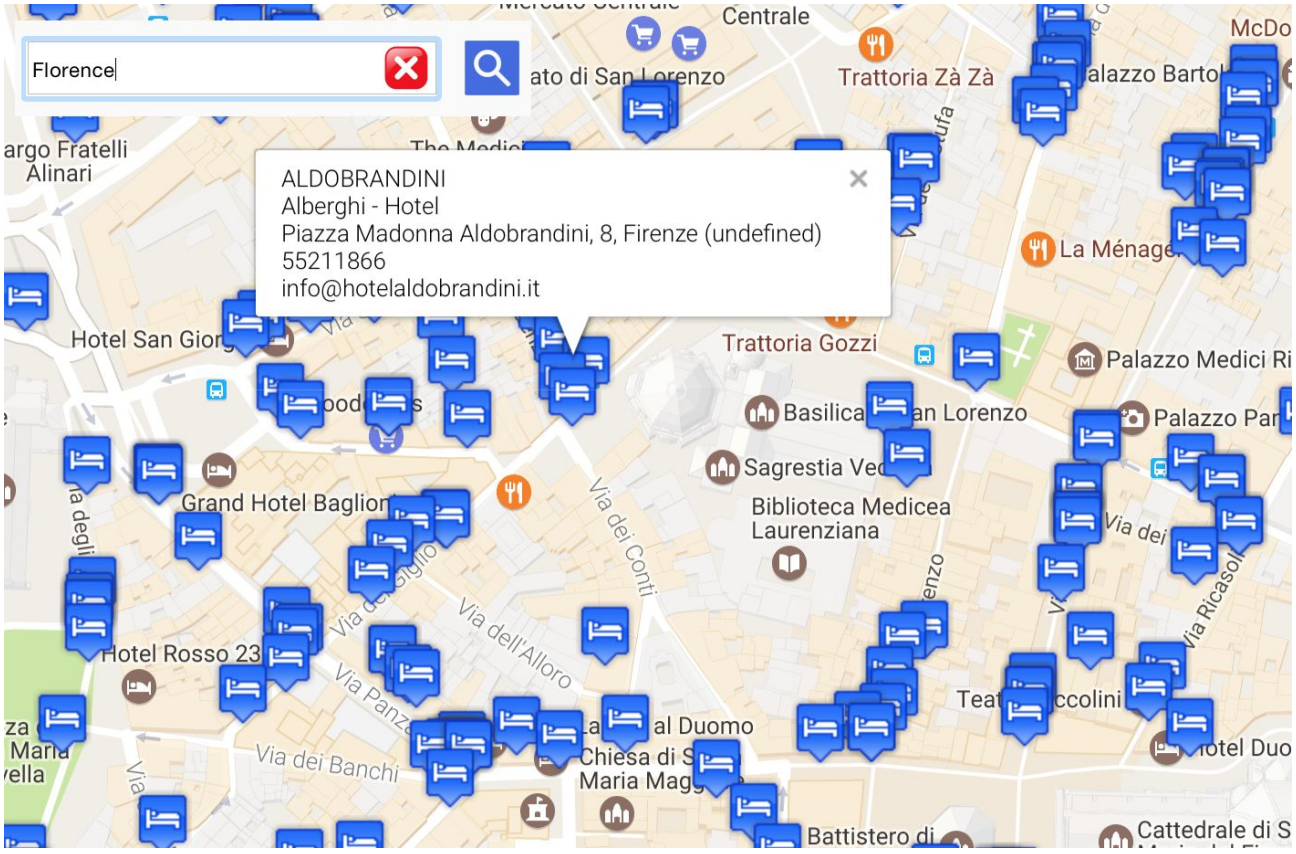

Fig. 7

5.5 License

All open data provided by Italian Regions are protected by a license (Table 6). Most datasets are provided under the Creative Commons License, the rest under the Italian Open Data License. Tourpedia provides the complete dataset under the most restrictive license used by Italian Regions. Considering that the CC0 1.0 Universal has no restrictions on its use and Italy Open Data License

[23] https://maps.google.com/

and CC-BY allow the same freedom, one of the two licenses could be chosen. Tourpedia provides its data under the CC-BY licence[24], being it more easily recognized at the international level.

Region	License
Basilicata	Not available
Emilia-Romagna	CC-BY-RER (CC-BY by Region Emilia-Romagna)
Friuli-Venezia Giulia	Italian Open Data License
Liguria	CC-BY
Lombardia	Italian Open Data License
Marche	Not available
Piemonte	CC0 1.0 Universal
Puglia	Italian Open Data License
Tuscany	CC-BY
Trentino-Alto Adige	CC-BY
Umbria	CC-BY
Veneto	CC-BY

Table 6. The type of data license for each Region.

[24] https://creativecommons.org/licenses/by/3.0/

5.6 Users

Tourpedia is envisaged for three types of users: *service providers*, *application developers* and *tourists*. A service provider is an entity, which shares data with tourists. Currently, Italian, French and Spanish Regions are considered to be service providers within Tourpedia. However, other categories of services providers can be imagined, such as museums, archives and so on. Often, open data are produced but not exploited by any application. Service providers may find in Tourpedia a place where their data are exploited. Thus, service providers could advertise their data by providing a link to Tourpedia. In addition Tourpedia may become a use case where errors and missing information in data are found. For example, the use of the Tourpedia Web application has permitted to discover some errors related to geographical coordinates of accommodations, located in Umbria. A mail has been sent to the Umbria Region to notify the problem and at the moment, this Region is trying to solve it.

Application developers are entities using the Web APIs provided by Tourpedia to build their mobile/desktop applications for tourists. Most of the Web APIs related to accommodations or in general to tourism are released under some restrictive terms. Tourpedia, instead, provides a full Web API, which is completely free. This constitutes an important aspect especially for the implementation of research prototypes. A community[25] for developers has been built to help them to use Tourpedia and the resources it offers.

Finally, tourists are the main beneficiaries of Tourpedia, because they exploit data contained in Tourpedia to organize their trips, select the best accommodations for their holidays and so on. In details, every traveller could exploit the Web application to search for accommodations located in a place and select his/her preferred one on the basis of its geographic position. At the moment, Tourpedia contains only accommodations but the idea consists in extending Tourpedia also to tourist attractions, events and other cultural information extracted from open data. In this way a tourist will be able to search for an accommodation located near a specific tourist attraction or a given event. The hope is that tourists will use Tourpedia also to build tours, involving different tourist attractions and events, as well as discover new unknown places, far from the most famous tourist tours.

6. Conclusions and Future Work

The work done in Tourpedia permitted the authors to deepen the world of open data. This presented both positive and negative sides. If, on the one hand, some very complete and frequently updated datasets have been found, on the other hand, it has been shown how some Regions do not provide data concerning accommodations and, in the case of Molise, it was not even possible to find a Web site containing open data.

[25] https://groups.google.com/forum/?tourpedia#!forum/tourpedia

6.1 Theoretical Implications

Technologies implemented by Tourpedia could have different implications for research, practice and society. Often, researchers encounter some difficulties to find open data about a specific topic, to perform their experiments and verify their theories. Even if data are available, they are provided in different formats, which often are incompatible, each other. Tourpedia gives researchers the possibility to use open data about tourism, which are levelled to a single format. Data can be accessed through the Web API freely and used to make experiments and verify theories.

From a practice point of view, the use of open data opens the road towards data liberalization. Although open data are not regulated through a standard, Tourpedia tries to define a common format for open data about tourism thus opening the road towards the definition of a standard about tourism data.

The current version of Tourpedia opens also the road to three new possible social scenarios. First of all, the current knowledge base could be transformed into an open source of updated information about tourism. This would constitute a valid alternative to proprietary Web sites, such as Booking.com and Trip Advisor. Secondly, the Web application could be empowered in order to support tours, which include some tourist attractions and/or events and suggest the nearest accommodations. Last but not least, in some way Tourpedia may help tourists to discover unknown places.

6.2 Practical Implications

The most important practical implication of Tourpedia concerns the construction of a single access point for all Italian, French and Spanish open data about accommodations. This constitutes a completely free knowledge base, which can be used directly by tourists (through the Web application) or by applications developers (through the Web APIs).

Another important implication of Tourpedia regards the construction of a simple mechanism for the integration of new sources of open data. Other categories of open data could be added easily to Tourpedia, with a minimum effort, such as tourist attractions and events. In addition, any source of data can ask to be added to Tourpedia, provided that it releases its data in a format compatible with Tourpedia.

6.3 Limitations and Future Research

Although a great effort was done to build the dataset, Tourpedia presents some limitations. First of all, extracted data are not homogeneous and often they are incomplete or wrong. This leads to the loss of data accuracy and in general reduces data quality. Unfortunately, this problem does not depend on Tourpedia, but it derives from a bad construction of the original dataset at the source. At the moment, the strategy adopted by Tourpedia to overcome this problem has consisted in trying to contact directly the source of data. However, this strategy has not been successful, because most sources have not answered to the sent notifications. Another possible strategy could involve the

accommodations ‘owners directly, by asking them to complete missing information or correct wrong ones. However, also in this case, the risk could be that only few accommodations ‘owners participated to the campaign. A mechanism based on rewards should be implemented to incentivize owners to give their data. For example, a dashboard could be provided to every accommodation’s owner with some statistics about the accommodation, such as the number of attractions and next events located nearby.

As further limitation, Tourpedia contains only accommodations. This limits the usage of the Web application by tourists, who can perform only a simple search about accommodations. However, the mechanisms implemented in Tourpedia are not domain dependent, thus they can be adapted easily to other domains, such as tourist attractions and events.

Another limitation of Tourpedia is the geographic coverage of the dataset. In fact, at the moment Tourpedia covers only some Italian, French and Spanish Regions. This limitation could be easily overcome by performing a search about accommodations datasets provided by other countries.

The work done in Tourpedia prepares the land for future developments, that will enable the implementation of new features and services. Tourpedia could play an important role within the tourism landscape, mainly thanks to its open features that will allow developers and researchers in the tourism sector to take advantage of a unified dataset. The development of Tourpedia is at the starting line, but has already made a good part of the work required to enter this world so crowded and full of competition.

List of Figures

References

Amichai-Hamburger, Y., McKenna, K. Y. and Tal, S. A. (2008), “E-empowerment: Empowerment by the Internet”, *Computers in Human Behavior*, Vol. 24 No. 5, pp. 1776-1789.

Bertot, J. C., Jaeger, P. T. and Grimes, J. M. (2010), “Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies”, *Government Information Quarterly*, Vol. 27 No. 3, pp. 264-271.

Chaves M.S., de Freitas, L.A. and Vieira, R. (2012), "Hontology: A multilingual ontology for the accommodation sector in the tourism industry", In Joaquim Filipe and Jan L. G. Dietz (Ed.), *KEOD*, SciTePress, pp. 149–154.

Cresci, S. D'Errico, A. Gazzè, D., Lo Duca, A., Marchetti, A. and Tesconi, M. (2014), "Towards a DBpedia of Tourism: the case of Tourpedia", in *Proceedings of the 2014 International Semantic Web Conference, Posters & Demonstrations Track-Volume 1272 in Riva del Garda, Italy 2014*, CEUR.org, pp. 129-132.

Dell'Erba, M., Fodor, O., Ricci, F. and Werthner, H. (2002), "Harmonise: A Solution for Data Interoperability", In Monteiro J.L., Swatman P.M.C., Tavares L.V. (Eds), *Towards the Knowledge Society. IFIP - The International Federation for Information Processing, vol 105,* Springer, Boston, MA, pp. 433-445.

Faizan, A., (2016), "Hotel website quality, perceived flow, customer satisfaction and purchase intention", *Journal of Hospitality and Tourism Technology*, Vol. 7 Issue: 2, pp.213-228.

García, A., Linaza, M. T., Franco, J., and Juaristi, M. (2015), "Methodology for the Publication of Linked Open Data from Small and Medium Size DMOs", In Tussyadiah, I. & Inversini, A. (Eds.) *Information and Communication Technologies in Tourism 2015*, Springer International Publishing, pp. 183-195.

Gazzè, D. Lo Duca, A. Marchetti, A. and Tesconi, M. (2015), "An overview of the Tourpedia linked dataset with a focus on relations discovery among places", in *Proceedings of the 11th International Conference on Semantic Systems in Vienna, Austria, 2015,* ACM, pp. 157-160.

Gibbs, C., Gretzel, U. and Saltzman, J. (2016), "An experience-based taxonomy of branded hotel mobile application features", *Information Technology & Tourism*, Vol. 16 No. 2, pp. 175-199.

Hepp. M. (2013), "Accommodation ontology language reference", Technical report, Hepp Research GmbH, Innsbruck.

Lakomaa, E., Kallberg and J., (2013), *Open Data as a Foundation for Innovation: The Enabling Effect of Free Public Sector Information for Entrepreneurs*, in *IEEE Access vol. 1, 2013*, pp. 558-563.

Legrand, B. (2004), "Semantic Web Methodologies and Tools for Intra-European Sustainable Tourism", White paper, Paris, Mondeca.

Li, J., Xu, L., Tang, L., Wang, S. and Ling, L. (2018), "Big data in tourism research: A literature review", *Tourism Management*,Vol. 68, pp. 301-323.

Longhi, C., Titz, J. B. and Viallis, L. (2014), "Open data: Challenges and opportunities for the tourism industry". In M. Mariani, R. Baggio, D. Buhalis, & C. Longhi (Eds.), *Tourism*

management, marketing, and development. Volume I: The importance of networks and ICTs, Palgrave Macmillan, New York, pp. 57-76.

Maccani, G., Donnellan, B. and Helfert, M. (2015), "Exploring the Factors that Influence the Diffusion of Open Data for New Service Development: an Interpretive Case Study", in *Proceedings of 23rd European Conference on Information Systems, Muenster, Germany*, 2015, paper 127.

Mariani, M. Di Felice, M. and Mura, M. (2016), "Facebook as a destination marketing tool: Evidence from Italian regional Destination Management Organizations", *Tourism Management*, Vol. 54, pp. 321-343.

McGuire, K. A. (2015), "Hotel Pricing in a Social World: Driving Value in the Digital Economy", John Wiley & Sons, North Carolina, USA.

McNaughton, M., McLeod, M.T. and Boxill, I., (2016), "An Actor Network Perspective of Tourism Open Data", in Metin Kozak , Nazmi Kozak (Ed.), *Tourism and Hospitality Management (Advances in Culture, Tourism and Hospitality Research, Volume 12)*, Emerald Group Publishing Limited, pp.47-60.

Moreno, M.C, Hörhager, G. Schuster, R. and Werthner, H. (2015), "Strategic E-Tourism Alternatives for Destinations. Information and Communication Technologies in Tourism 2015", in Tussyadiah I., Inversini A. (Eds) *Information and Communication Technologies in Tourism 2015*, Springer, Cham, pp. 405-417.

Pantano, E., Priporas, C. and Stylos, N. (2017), "You will like it! using open data to predict tourists response to a tourist attraction", *Tourism Management*, Vol. 60, pp. 430-438.

Pesonen, J. and Lampi, M. (2016), "Utilizing open data in tourism", In Tussyadiah, I. & Inversini, A. (Eds.) *Information and Communication Technologies in Tourism 2016*, Springer International Publishing, pp. 1-5.

Prantner K, Ding Y, Luger M and Yan Z (2007), "Tourism ontology and semantic management system: state-of-the-arts analysis", in *Proceedings of IADIS (International Association for Development of the Information Society) international conference WWW/Internet 2007. Vila Real, Portugal, 2007*, p. 111-15.

Scarinci, J. and Myers, T. (2014), "A Semantic Web framework to enable sustainable lodging best management practices in the USA", *Information Technology & Tourism*, Vol. 14 No. 4, pp. 291-315.

Soualah-Alila, F. Coustaty, M. Rempulski, N. and Doucet, A. (2016), "DataTourism: Designing an Architecture to Process Tourism Data", In Tussyadiah, I. & Inversini, A. (Eds.), *Information and Communication Technologies in Tourism 2016*, Springer International Publishing, pp. 751-764.

UNWTO (2017) “World Tourism Barometer, vol.15 – June”, available at: http://cf.cdn.unwto.org/sites/all/files/pdf/unwto_barom17_03_june_excerpt_1.pdf (accessed 2018/05/30).

Urata, M., Ogishima, K., Fukuyasu, M., Endo, M. and Yasuda, T. (2016), “Promotion of local government open data for sightseeing events”, *Journal of Global Tourism Research*, Vol. 1 No. 2, pp. 133-138.

Varkaris, E. and Neuhofer, B. (2017), “The influence of social media on the consumers’ hotel decision journey”, *Journal of Hospitality and Tourism Technology*, Vol. 8 Issue: 1, pp.101-118.

Wu, C. T., Liu, S. C., Chu, C. F., Chu, Y. P., and Yu, S. S. (2014), “A study of open data for tourism service”, *International Journal of Electronic Business Management*, Vol. 12 No. 3, pp. 214-221.

Zeng, B. and Gerritsen, R. (2014), “What do we know about social media in tourism? A review”, *Tourism Management Perspectives*, Vol. 10, pp. 27-36.